分段提取函数型数据特征的算法研究

金海波 [1]，马海强 [2]

(1. 太原科技大学 数学系，太原 030024; 2. 江西财经大学 统计学院，南昌 330013)

摘　要：针对函数型数据分类算法中全局统计特征表达能力有限，且显著点特征易受噪声干扰等问题，提出一种基于统计深度方法的函数曲线特征分段提取算法。首先，利用数据平滑技术对离散观测的数据进行平滑化处理，同时引入函数型数据的一阶和二阶导函数；然后，分段计算函数本身及其低阶导函数的马氏积分深度值，在此基础上构造函数曲线特征向量；最后，给出三种选择调节参数的搜索方案，并进行分类研究。在 UCR 数据集上的实验表明，与当前其他曲线特征提取算法相比，所提算法能有效提取函数曲线特征，提高分类的准确性。

关键词：函数型数据；分段特征；深度函数；函数型数据分类

中图分类号：TP301.6　　doi: 10.19734/j.issn.1001-3695.2018.11.0873

Segmental feature extraction for functional data

Jin Haibo[1], Ma Haiqiang[2]

(1. *Dept. of Mathematic*, *Taiyuan University of Science & Technology*, *Taiyuan* 030024, *China*; 2. *School of Statistics*, *Jiangxi University of Finance & Economics*, *Nanchang* 330013, *China*)

Abstract: Since the representation ability of statistical global feature for functional data classification is limited, and the salient point feature is susceptible to noise disturbance, proposed a segmental feature extraction algorithm based on statistical depth notion. Firstly, the smoothing technique is used to pre-smooth the discrete observed data, and the first and second derivatives of the function curves are defined accordingly. Then, depths of Mahalanobis integral of the functions and its low-order derivatives in segments are calculated, and thus feature vectors of function curves are constructed based on the depth measures. Finally, the optimal number of segments for classification is selected by data-driven, and the binary classification of function data is studied. Compared with the other curve feature extraction algorithms, experiments on UCR datasets show that the proposed algorithm performs well in extracting the feature of curve, and improves the classification accuracy effectively.

Key words: functional data; segmental feature; depth function; functional data classification

0　引言

近三十年来，随着科学技术的迅猛发展，人们获取和存储数据的能力得到了极大提高。在现实生活的很多领域中，人们越来越多需要处理具有实时性、空间型等函数特性的数据，如经济活动中的金融数据、工业设备产生的传感器数据、环境科学中的气象数据等。这些数据往往是带噪声的离散观测数据，在实际数据的分析中需要重新有效表达这些序列数据，再根据研究或应用目的，选用合适的数据挖掘技术进行分析，如分类或聚类分析。其中，若将序列数据看成函数型数据[1]，便可以充分利用函数的优良性质和特点，极大地提升数据挖掘的深度和精度。

函数型数据本质上是一类无穷维数据，其样本单元是无穷维函数空间中的随机曲线，即为随机过程的一次实现。函数型数据分析方法主要采用泛函分析中的相关方法对函数型数据进行建模，它非常侧重数据的函数特性，即通常将一定时间范围内的观测数据看成一个整体，而不对数据内部的相依性质设置任何假定。针对不同观测个体，函数型数据分析方法允许使用不同抽样技术在不同时间点上获得稀疏或稠密的观测值。

值得注意的是，函数型数据内在的无穷维特性将会不可避免地给实际数据分析带来诸多挑战。如果直接在函数型数据上进行数据挖掘，相关数值计算工作量非常大[2]，因此，对函数型数据进行降维处理或特征提取是必要的[3~5]，获得低维特征后，后续便可采用成熟的数据挖掘技术进行分析，从而提高计算效率。特别值得指出的一点是，根据特定的分析目的和应用领域，应该对函数型数据采取不同的特征提取方法。例如针对分类问题，为了提升分类精度，应该提取与类别相关的数据特征[4,5]。

国内外很多专家学者对函数型数据的分类问题进行了大量研究。考虑到特征提取和分类算法之间的紧密联系，特别是特征提取的优劣需要用分类结果来评价，因此，在介绍前人工作时，将对文献中分类算法和采用的特征提取技术一并介绍。Alonso 等人[6]根据函数曲线和类均值曲线间距离重新构造判别变量，进而采用线性判别分析 LDA 或最近邻 KNN 等多元分类技术进行分类；除函数本身外，算法还利用了多阶导函数去构造判别变量。实验表明，借助一阶和二阶导函数构造的判别变量可以显著降低错分率。Torrecilla 等人[7]提出一种极大值搜索的函数特征选择迭代算法（RMH），他们首先通过计算随机函数 $X(t), t\in[0,T]$ 和类变量 Y 的距离相关系数，得到最大系数值对应的 t_0 显著点；然后在子区间 $[0,t_0)$ 和 $(t_0,T]$ 上递归搜索去掉 $X(t_0)$ 影响后的显著点；最后利用 $X(t)$

收稿日期：2018-11-28；**修回日期**：2019-01-22

作者简介：金海波（1980-），男，山西闻喜人，讲师，硕士，主要研究方向为数据挖掘（jhb800@qq.com）；马海强（1982-），男，山西晋城人，讲师，硕导，博士，主要研究方向为函数型数据分析、分位数回归等.

降维得到的点集 $\{X(t_i)\}_{i=0}^{p}$ 进行最近邻分类。Dai 等人[8]提出一种基于似然比的贝叶斯分类算法，并从理论上证明了其“完美分类”的性质。他们首先对函数型数据进行投影变换，得到多个独立的主成分得分；然后分别对这些主成分得分的概率密度函数进行非参估计；最后利用似然比公式完成贝叶斯分类。Mosler 等人[9]通过两步变换方法，即首先把函数和一阶导函数的分段积分值作为特征向量，再应用多元深度函数把特征向量映射到二维值空间 DD-plot；最后运用最近邻分类和 DDα 过程进行分类。Li 等人[10]首先使用 F 统计量求得曲线显著点及其相邻子区间，然后利用 LDA 提取曲线特征，最后运用支持向量机进行分类。此方法适用于空间异质或不规则抽样的曲线数据。Fraiman 等人[11]利用一组函数来定义曲线特征，并分别应用在分类、回归和主成分分析等方面。Rossi 等人[12] 则详细介绍了支持向量机在函数型数据分类中的作用。

国内学者马忱等人[13]提出了面向函数型数据的结合主成分分析法和最小凸包法的快速特征选择(FFS)方法,他们所提方法不仅可以快速获得稳定的特征子集，而且具有很好的实际效果。苏本跃等人[14]则利用函数型数据分析方法，将可穿戴式运动捕捉系统采集的人体周期行为数据进行函数化处理，准确地定义了数据的连续性与周期性，最后，根据不同行为一个周期内的曲线特征差异，利用支持向量机对动态行为进行分类识别。

1 分段特征提取(SFE)算法

考虑到函数对象的全局统计特征表达能力有限，且显著点等局部特征易受噪声干扰，本文提出了基于统计深度函数的分段特征提取算法（SFE）。鉴于函数型数据平滑特性，所提算法不仅利用了函数对象本身的特点，而且还利用多阶导函数的分段特征，因此，所提算法可以全面刻画函数型数据的变化特征。UCR 多个数据集的实验验证了所提算法在函数型数据的分类应用上具有很好的实际效果。

1.1 问题定义与算法流程

为描述方便，考虑函数型数据二分类问题。设 $X(t), t\in[0,T]$ 是来自概率空间 (Ω, F, P) 的时间连续随机过程，函数对象 $X_i(t)$ 是此过程的一次实现（或轨迹），$\{X_i(t), Y_i\}_{i=1}^{N}, t\in[0,T]$，是一组由函数和类别标签组成的数据对集合，其中 $Y_i\in\{0,1\}$ 是分类变量，这些函数（轨迹）来自两个不同的总体,类别 $Y_i=0$ 代表 P_0，类别 $Y_i=1$ 代表 P_1。分类问题是对未知类别的函数对象 $X^{test}(t)$ 推断所属的总体 P_0 或 P_1。本质上，特征提取就是要找到一种形如 $\Phi: F\rightarrow R^d$ 的映射，使得对无穷维函数型数据的分类问题可以转换为 R^d 中的有限维分类问题，从而避免无限维函数空间下分类不可行问题。

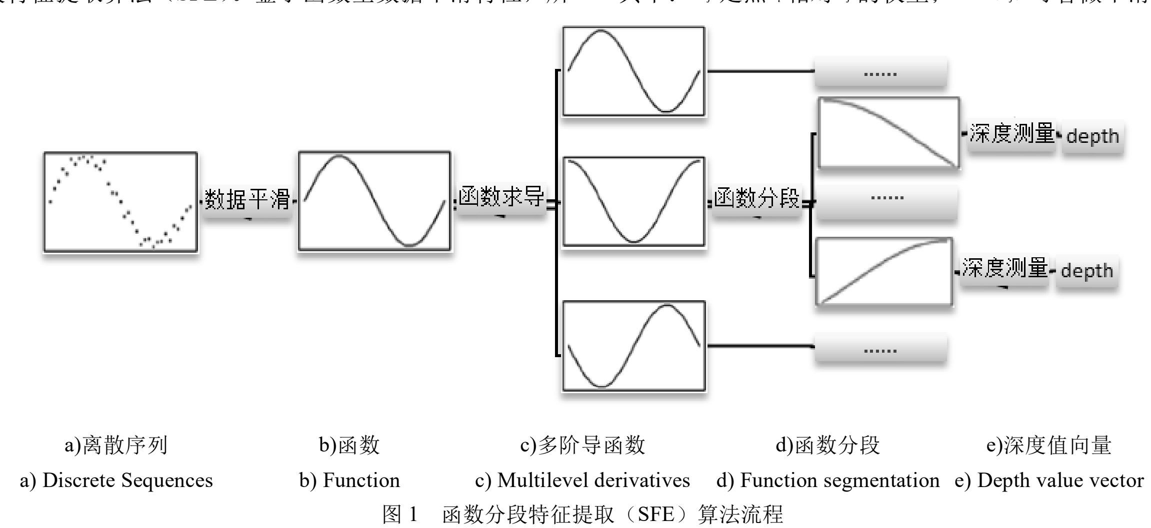
图 1 描述了 SFE 算法处理数据的主要流程。图 1(a)为含噪声的离散序列数据，经数据平滑后得到连续的函数型数据，如(b)所示；然后求取其低阶导函数，如(c)所示，从上到下依次为原函数、一阶和二阶导函数；再对这三类函数分段，如(d)所示；最后计算每个分段的统计深度值，并组合得到深度值向量，如(e)所示，即得到函数的特征向量。

1.2 数据平滑

如果待研究的观测样本是含噪声的数据序列，$Y(t_i), i\in\{1,\ldots,m\}$，即满足模型 $Y(t)=X(t)+\varepsilon(t)$，其中残差 $\varepsilon(t)$ 独立于 $X(t)$，则可用线性平滑方法得到原始 $X(t_i)$，即

$$\hat{X}(t_i)=\sum_{j=1}^{m} s_{ij}Y(t_i)\ , \qquad (1)$$

其中：s_{ij} 是点 t_i 相对 t_j 的权重；$S=(s_{ij})$ 可看做平滑矩阵。



a)离散序列 b)函数 c)多阶导函数 d)函数分段 e)深度值向量

a) Discrete Sequences b) Function c) Multilevel derivatives d) Function segmentation e) Depth value vector

图 1 函数分段特征提取（SFE）算法流程

Fig. 1 Procedure of segmental feature extraction for functional data

目前，主要有两种线性平滑方法用来恢复原始 $X(t)$。一种方法是利用一组基函数 $\{\Phi_k\}_{k\in N}$ 的线性组合来近似逼近 $X(t)$，这里选择足够多的 k_n 个基函数，即

$$X(t)=\sum_{k\in N} c_k\Phi_k(t)\approx\sum_{1}^{k_n} c_k\Phi_k(t) \qquad (2)$$

在本文所提算法中，使用一组 B 样条基函数来逼近原函数。

另一种方法则是采用非参数核平滑技术[2]，使用 Nadaraya-Watson 估计平滑矩阵 $S=(s_{ij})$：

$$s_j(t_i)=K(\frac{t_i-t_j}{h})\Big/\sum_{k=1}^{m}K(\frac{t_i-t_k}{h}) \qquad (3)$$

其中：$k(\bullet)$ 为核函数，一般选用高斯核函数。最佳窗宽参数 h 可通过交叉验证方法获得。

为了比较两种平滑方法的差异，图 2 给出了 B 样条曲线平滑和非参核平滑的示意图。图中曲线数据来自于 GunPoint 数据集中第 34 条曲线片段，其中共包含 21 个数据点。B 样条基函数个数和核函数窗宽参数是从一组可选值中通过交叉验证计算得到。就此例而言，从图 2 可以看出，第一种平滑方法效果比第二种方法好。

1.3 统计深度函数

统计深度函数和相关分位数函数可以对多元数据进行非参数描述和结构分析，根据不同的中心性概念和定义，存在多种类型的深度函数。对函数型数据，也可以定义深度函数，用来刻画某条曲线(过程)相对曲线样本的向心性度量[15,16]。

考虑曲线 X_i，来自样本 $P=\{X_i(t)\}_{i=1}^n, t\in[0,T]$,本文算法使用如下定义的马氏(Mahalanobis)积分深度函数[16]：

$$FMD(X_i,P)=\int_0^T(1-|1/2-F_{n,t}(X_i(t))|)\,dt \tag{4}$$

$$F_{n,t}(X_i(t))=n^{-1}\sum_{k=1}^n I(X_k(t)\le X_i(t)) \tag{5}$$

其中：式(4)中被积函数表示一维的深度函数；式(5)中 $I(\cdot)$ 表示示性函数。

式(4)本质上可以看成是深度函数在函数型数据的一种推广。类似于一维随机变量中的次序统计量，马氏积分深度函数主要用来刻画出一条曲线在整个数据中所处的位置信息。马氏积分深度函数不仅计算简单，而且具有很好的稳健性质。此外，基于式(4)，本文还可以用来构造诸多函数型数据的统计量，如函数型数据的秩和截断均值等，从而可以克服数据中异常点的影响，得到更精确可信的分析结果。

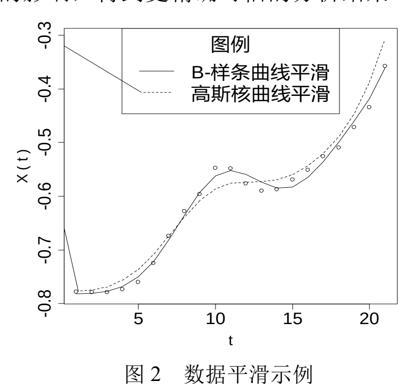


图 2　数据平滑示例

Fig. 2　Example of data smoothing

1.4　分段特征提取

假设 $X^{(1)}$ 和 $X^{(2)}$ 是来自总体 $P_k, k\in\{0,1\}$ 的函数样本，任取 $X_i\in X^{(1)}\cup X^{(2)}$，函数 X_i 连续且光滑，D^0X_i、D^1X_i 和 D^2X_i 分别表示函数本身、一阶和二阶导函数，这三类函数分别描述了函数曲线的位置、斜率变化和凹凸性质。若考虑如下变换 $\Phi:F\to R^6$，即

$$X_i\mapsto[\,f(D^0X_i,X^{(1)}),\ f(D^0X_i,X^{(2)}),\ f(D^1X_i,D^1X^{(1)}),\ f(D^1X_i,D^1X^{(2)}),\ f(D^2X_i,D^2X^{(1)}),\ f(D^2X_i,D^2X^{(2)})\,] \tag{6}$$

其中：$f(D^pX_i, D^pX^{(k)})$ 是实值函数映射，表示函数 D^pX_i 相对样本 $D^pX^{(k)}$ 的统计深度值，这里应用式(4)和(5)定义的积分深度函数。注意到上其中 f 变换针对 D^pX_i 定义域 $[0,T]$ 进行，可得到函数 D^pX_i 类别相关的全局统计特征。为了提高 D^pX_i 的局部特征表达能力，对 D^pX_i 分段应用 f 变换。为描述方便，这里讨论 D^0X_i 即 X_i 的分段变换，D^1X_i 和 D^2X_i 可同理分析。考虑将定义域 $[0,T]$ 分成 N_l 个等距子区间 $[0,T/N_l)$，$[T/N_l, 2T/N_l),\cdots$，$[(N_l-1)T/N_l, T)$

每个子区间上的函数分段用 $X_{i,j}, j=1,2,\ldots N_l$ 表示，再对每个 $X_{i,j}$ 实施 f 变换；同理，对 D^1X_i 和 D^2X_i 划分成 N_s 和 N_c 个分段，满足 $N_l+N_s+N_c\ge 1$。综上，求得函数空间到特征空间的变换为 $\Phi:F\to R^{2(N_l+N_s+N_c)}$，即

$$X_i\mapsto[\,f(D^0X_{i,0},X_0^{(1)}),\cdots,f(D^0X_{i,N_l},X_{N_l}^{(1)}),$$
$$f(D^0X_{i,0},X_0^{(2)}),\cdots,f(D^0X_{i,N_l},X_{N_l}^{(2)}),$$
$$f(D^1X_{i,0},D^1X_0^{(1)}),\cdots,f(D^1X_{i,N_s},D^1X_{N_s}^{(1)}),$$
$$f(D^1X_{i,0},D^1X_0^{(2)}),\ldots,f(D^1X_{i,N_s},D^1X_{N_s}^{(2)}),$$
$$f(D^2X_{i,0},D^2X_0^{(1)}),\ldots,f(D^2X_{i,N_c},D^2X_{N_c}^{(1)}),$$
$$f(D^2X_{i,0},D^2X_0^{(2)}),\ldots,f(D^2X_{i,N_c},D^2X_{N_c}^{(2)})\,] \tag{7}$$

1.5　调节参数的选择

SFE 算法中三个最重要的参数是分段数 N_l、N_s 和 N_c，这些参数值的选择将直接影响后续分类算法性能，正确选择这三个参数仍缺乏理论依据[9]。针对不同数据集的实验表明，通常选取不大于 10 的分段数就可以取得满意结果，但穷尽搜索三个参数的最佳组合仍然花费很长时间，为此需要设计启发式搜索策略来提高效率。这里考虑两种简化方案，第一种是独立选择法，即单独搜索三个参数并确定各自的最佳分段数为 n_l、n_s 和 n_c，这种方法实现简单，搜索任务可以并行化处理；第二种方案是分步选择法，首先假设 (N_s,N_c) 为 $(0,0)$，搜索并确定 N_l 的最佳分段数是 n_l，然后固定 (N_l,N_c) 值为 $(n_l,0)$，在此条件下搜索 N_s 最佳分段数为 n_s，最后固定 (N_l,N_s) 值为 (n_l,n_s)，搜索得到 N_c 最佳分段数为 n_c，最终取得参数 N_l、N_s 和 N_c 的最佳组合为 (n_l, n_s, n_c)。易见，独立选择法和分步选择法搜索效率远高于穷尽搜索方法。

1.6　算法伪代码

综上讨论，基于深度信息的函数特征分段提取算法 SFE 伪代码如算法 1 所示。为简单起见，只列出核心部分代码。这里对算法 1 做必要解释，X 和 Y 分别表示函数样本集和相应的类别，N_l, N_s 和 N_c 分别表示函数及其一阶和二阶导函数的分段数目。代码第 1~7 行计算函数分段的统计深度值，第 3 行把样本集按类别分为两类子集，第 5、 6 行根据式 (4) 分别计算深度值。第 8~14 行对函数进行分段并计算其分段特征值，其中第 10 行对函数进行分段，第 11~13 行循环调用 depth 函数计算各分段的统计深度值。第 15~23 行是算法主过程，第 16 行应用式 (2) 平滑原始观测数据，第 17、18 行分别获得函数一阶和二阶导函数，第 19~21 行计算函数及其导函数的分段特征值，第 22 行把函数及导函数的特征值组合成特征矩阵。

算法 1　分段特征提取算法

```
输入: X, Y, Nₗ, Nₛ, N꜀ 。
输出: features (特征矩阵) 。
1  function  depth(Xᵢ, Y)
2      d ← matrix(size(Xᵢ), 2)
3      {Xᵢ(1), Xᵢ(2)} ← groupBy(Xᵢ , Y)
4      for(k in 1:size(Xᵢ)) {
5          d(k,1) ← FMD(Xᵢ[k], Xᵢ(1))
6          d(k,2) ← FMD(Xᵢ[k], Xᵢ(2)) }
7      return d
8  function  seg_features(X, Y, n)
9      fx ← [ ]
10     {X₁ , … , Xₙ} ← segment(X, n)
11     for(i in 1:n) {
12         d ← depth(Xᵢ, Y)
13         fx ← [fx  d] }
14     return fx
15  procedure SFE(X, Y, Nₗ, Nₛ, N꜀)
16     D⁰X ← smoothing(X, Y)
17     D¹X ← deriv(D⁰X, 1)
18     D²X ← deriv(D⁰X, 2)
19     f_D⁰X ← seg_features(D⁰X, Y, Nₗ)
20     f_D¹X ← seg_features(D¹X, Y, Nₛ)
```

```
    f_D²X ← seg_features(D²X, Y, N_c)
    features ← [f_D⁰X  f_D¹X  f_D²X]
    return features
```

2　实验及结果分析

2.1　数据集介绍

为了比较不同算法下的分类性能，本文选取了 UCR 标准序列数据集[17]中的六个数据集进行实验。这些数据曲线特征复杂，对其分类具有极大的挑战性。每个数据集均包含独立的训练集和测试集，详细描述如表 1 所示。图 3 画出了 WormsTwoClass 数据集中的两类曲线片段样例。

表 1　实验数据集

Table 1　Experimental datasets

数据集	训练数	测试数	序列长度
GunPoint	50	150	150
BeetleFly	20	20	512
Ham	109	105	431
Herring	64	64	512
Earthquakes	139	322	512
WormsTwoClass	77	181	900

2.2　实验设计

本文提出一种函数型数据特征的分段提取算法 SFE，得到低维数据特征后，再采用成熟算法进行分类。为了验证本文算法的通用性，实验采用 SFE 算法和分类算法的多种组合方案。LDA 表示线性判别分析方法，SVM 表示采用径向基核函数支持向量机方法，其中参数 c = 8，gamma = 0.5。RF 表示随机森林分类方法，其分类器参数都采用默认值。为便于比较，各分类算法的参数值在不同数据集上运行均未做任何优化。

算法代码使用 R 语言实现，第三方工具包主要包括函数型数据分析包 fda.usc 和分类回归训练包 caret，其中后者用来构建分类模型，包括模型训练和分类预测两个过程。使用训练集拟合模型和参数寻优，采用 K 折交叉验证模型，依据训练集规模 K 可取 5 或 10。模型建立后，使用独立测试集评估模型性能，算法选用分类精度作为评价指标。

2.3　结果及分析

表 2 列出了不同分类方法下数据集上的分类精度。前两列是两种简单非参分类模型的分类结果[17]，它们作为评价本文算法性能的基准数据，其中 1NN 表示以欧氏距离为相似性度量的最近邻分类，1NN-DTW 表示采用动态时间弯曲距离的最近邻分类。若曲线样本非时间对齐，1NN-DTW 分类比单纯用 1NN 分类结果好，如 Earthquakes 和 WormsTwoClass 数据集。表中后八列给出四种分类算法在曲线分段和不分段条件下的分类精度，其中不分段表示 SFE 算法参数 N_l、N_s 和 N_c 取值为 1，分段条件下分类精度很大程度上依赖于分段数目，表中给出最佳分段数 (n_l,n_s,n_c) 下的分类结果。需要指出的是，根据分段数参数搜索策略的不同，得到的分段数目并不唯一，可能会造成分类结果差异。

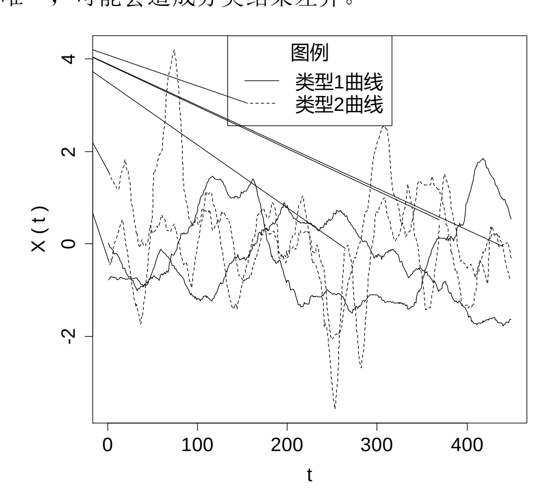


图 3　原始数据曲线样例

Fig. 3　Samples of original data curve

表 2　不同分类方法下的分类精度比较

Table 2　Classification accuracy comparison of different classification methods

数据集	1NN	1NN-DTW	SFE + LDA		SFE + SVM		SFE + RF	
			未分段	分段	未分段	分段	未分段	分段
GunPoint	0.913	0.907	0.880	0.953 (10,6,5)	0.893	0.947 (6,6,1)	0.920	0.953 (10,5,1)
Beetlefly	0.750	0.700	0.850	0.900 (5,4,6)	0.700	0.850 (5,1,3)	0.700	0.900 (10,7,2)
Ham	0.600	0.467	0.781	0.790 (7,2,1)	0.781	0.705 (7,5,1)	0.720	0.752 (7,5,10)
Herring	0.516	0.531	0.516	0.609 (10,5,0)	0.500	0.594 (3,2,4)	0.469	0.609 (3,4,1)
Earthquakes	0.674	0.742	0.798	0.814 (1,0,3)	0.786	0.817 (2,10,2)	0.801	0.811 (8,6,3)
WormsTwoClass	0.586	0.663	0.575	0.600 (7,4,3)	0.702	0.729 (7,3,6)	0.663	0.702 (7,3,6)

对表 2 中结果进行分析，得到如下结论：

a) 在曲线未分段情况下，采用 SFE 算法得到的分类结果在多数情况下和基准结果相差不大，说明统计深度值作为曲线特征是非常有效的。从 Ham 和 Earthquakes 数据集上结果来看，相较 1NN 基准结果，三种分类算法的分类精度平均提升 16.1%和 12.1%。

b) 在曲线分段情况下，从曲线得到的特征维数更多，所有分类算法下的分类精度高于未分段情况下的分类精度。这点在 GunPoint、BeetleFly 和 Herring 数据集上表现明显，三种分类算法下的分类精度相比未分段情况下平均提升 5.33%、13.3%和 10.9%。另外在 GunPoint 和 WormsTwoClasss 数据集上最佳分类精度分别是 95.3%和 72.9%，相比较文献[8]中相同数据集上最好分类结果分别提升 23.3%和 14.9%。

c) 为公平比较多个分类算法在不同数据集上的表现，并且突出 SFE 算法的作用，实验中未优化任何分类算法参数，未对特征过多预处理，可能出现非预期分类结果。比如 Ham 数据集上的 SVM 分类结果，由于过拟合问题，使得测试集分类精度不如预期。在实际数据分析中，可以采取特征预处理技术和分类算法参数优化工作便可避免上述问题。

2.4　与其他特征提取方法的比较

在对函数型数据进行特征提取时，常用降维方法有主成分分析法（PCA）[1]和偏最小二乘法（PLS）[4,5]。在文献[6]中作者提出了 DFM（DISTANCE TO THE FUNCTIONAL Mean）方法，其主要思想是根据如下定义提取函数及其导函数的特征，构造判别变量：

$$d=\left(\int_0^T\left|\left(X(t)-\overline{X(t)^{(1)}}\right)\right|^p dt\right)^{1/p}-\left(\int_0^T\left|\left(X(t)-\overline{X(t)^{(2)}}\right)\right|^p dt\right)^{1/p}\quad(8)$$

其中：d 表示提取到的实值特征；$X(t),t\in[0,T]$ 表示函数曲

线样例；$\overline{X(t)}^{(1)}$ 和 $\overline{X(t)}^{(2)}$ 表示正类和反类的类均值函数曲线；p 一般取 1 或 2。

为了比较多种特征提取方法的性能，本文选用文中介绍的六个数据集进行实验，后续分类算法均采用 LDA 方法，算法代码使用 R 语言实现。需要指出的是，PCA 方法根据方差累积贡献率来确定主成分个数，PLS 方法通过交叉验证方法获得成分个数，DFM 方法根据式（8）分别提取原函数、一阶和二阶导函数特征。所有方法均对原始数据进行函数化表达和平滑处理，在测试数据集上得到的分类精度如表 3 所示。

表 3　不同特征提取方法下的分类精度比较

Table 3　Classification accuracy comparison of different feature extraction methods

数据集	DFM	PCA	PLS
GunPoint	0.807	0.740	0.740
Beetlefly	0.750	0.650	0.800
Ham	**0.800**	0.667	0.695
Herring	0.578	**0.625**	0.594
Earthquakes	0.795	0.795	**0.820**
WormsTwoClass	0.558	0.536	0.586

结合表 3 中 LDA 分类结果，并与表 2 所得结果进行比较可得：在总共 18 个分析结果中，只有三种情形比本文方法分类精度高，分别是 Ham 数据集上的 DFM 方法、Herring 数据集上的 PCA 方法和 Earthquakes 数据集上的 PLS 方法。由此可见，本文所提的 SFE 方法整体上优于其他三种特征提取方法。另外，由于 PLS 方法考虑了样本数据和类变量的相关性，提取的特征质量更高，分类效果比 PCA 更好，但 PLS 方法和 DFM 方法各有优劣。

3　结束语

分类问题是函数型数据分析领域中的重要研究方向，能否有效提取函数型数据的低维特征非常关键。本文所提算法对函数及导函数曲线分段处理，基于统计深度方法，把无穷维函数变换为低维特征向量，再采用标准分类算法处理，从而避免了全局特征和显著点特征表达的不足，在多个数据集上的实验结果验证了文中所提 SFE 算法的有效性。进一步考虑如下三个问题: a)如何处理非时间对齐的样本曲线，如界标法校准等，将极大改善后续函数化表达及分析;b)当前算法中函数分段区间是等距的，能否提出启发式策略，自适应地确定非等距子区间，以便提取更具辨别的类特征; c)函数特征映射可以考虑更多变换形式，文中使用的统计深度值即向心性度量可拓展成多种定义，以上三点是下一步的工作重点。

参考文献：

[1] Ramsay J O, Silverman B W. Functional data analysis [M]. 2nd ed. New York: Springer, 2005.

[2] Ferraty F, Vieu P. Nonparametric functional data analysis: theory and practice [M]. New York: Springer, 2006.

[3] 孟银凤，梁吉业. 函数型数据分类中的稳健主成分分析 [J]. 小型微型计算机系统, 2016, 37 (7): 1499-1503. (Meng Yinfeng, Liang Jiye. Robust principal analysis in the classification for functional data [J]. Journal of Chinese Computer Systems, 2016, 37 (7): 1499-1503.)

[4] Preda C, Saporta G, Lévéder C. PLS classification of functional data [J]. Computational Statistics, 2007, 22 (2): 223-235.

[5] Delaigle A, Hall P. Methodology and theory for partial least squares applied to functional data [J]. Annals of Statistics, 2012, 40 (2012): 322-352.

[6] Alonso A M, Casado D, Romo J. Supervised classification for functional data: a weighted distance approach [J]. Computational Statistics and Data Analysis, 2012, 56 (7): 2334-2346.

[7] Torrecilla, José L, Suárez *et al.* Feature selection in functional data classification with recursive maxima hunting [C]// Advances in Neural Information Processing Systems. 2016: 4835-4843.

[8] Dai X, Myuller H G, Yao F. Optimal bayes classifiers for functional data and density ratios [J]. Biometrika, 2017, 104 (3): 545-560.

[9] Mozharovskyi P, Mosler K. Fast DD-classification of functional data [J]. Statistical Papers, 2017, 58 (4): 1055-1089.

[10] Li B, Yu Q. Classification of functional data: a segmentation approach [J]. Computational Statistics and Data Analysis, 2008, 52 (10): 4790-4800.

[11] Fraiman R, Gimenez Y, Svarc M. Feature selection for functional data [J]. Journal of Multivariate Analysis, 2016, 146: 191-208.

[12] Rossi F, Villa N. Support vector machine for functional data classification [J]. Neurocomputing, 2006, 69 (7-9): 730-742.

[13] 马忱，王文剑，姜高霞. 面向函数型数据的快速特征选择方法 [J]. 模式识别和人工智能, 2017, 30 (9): 822-832. (Ma Chen, Wang Wenjian, Jiang Gaoxia. Fast feature selection for functional data [J]. Pattern Recognition and Artificial Intelligence, 2017, 30 (9): 822-832.)

[14] 苏本跃，蒋京，汤庆丰，等. 基于函数型数据分析方法的人体动态行为识别 [J]. 自动化学报, 2017, 43 (5): 866-876. (Su Benyue, Jiang Jing, Tang Qingfeng *et al.* Human dynamic action recognition based on functional data analysis [J]. Acta Automatica Sinica, 2017, 43 (5): 866-876.)

[15] Sguera C, Galeano P, Lillo R. Spatial depth-based classification for functional data [J]. Test, 2014, 23 (4): 725-750.

[16] Serfling R, Wijesuriya U. Depth-based nonparametric description of functional data, with emphasis on use of spatial depth [J]. Computational Statistics and Data Analysis, 2017, 105: 24-45.

[17] Chen Y, Keogh E, Bing H *et al.* The UCR time series classification archive [DB/OL]. (2015) [2018-11-01]. http://www. cs. ucr. edu/~eamonn/time_ series_data.